ULTRA
SONO
GRAPHY

Automated versus handheld breast ultrasound examinations of suspicious breast masses: methodological errors in the reliability analysis

Siamak Sabour[1,2]

[1]Department of Clinical Epidemiology, School of Health and Safety, [2]Safety Promotions and Injury Prevention Research Centre, Shahid Beheshti University of Medical Sciences, Tehran, Iran

LETTER

https://doi.org/10.14366/usg.19038
pISSN: 2288-5919 • eISSN: 2288-5943
Ultrasonography 2020;39:102-103

Received: July 2, 2019
Revised: July 2, 2019
Accepted: August 1, 2019

Correspondence to:
Siamak Sabour, MD, MSc, DSc, PhD, Department of Clinical Epidemiology, School of Health and Safety, Shahid Beheshti University of Medical Sciences, Chamran Highway, Velenjak, Daneshjoo Blvd, Tehran 198353-5511, Iran

Tel. +98-21-22421814
E-mail: s.sabour@sbmu.ac.ir

How to cite this article:
Sabour S. Automated versus handheld breast ultrasound examinations of suspicious breast masses: methodological errors in the reliability analysis. Ultrasonography. 2020 Jan;39(1):102-103.

I was interested to read the recently published article by Yun et al. in the July 2019 issue of *Ultrasonography* [1]. The purpose of the authors was to assess the reliability of automated breast ultrasound (ABUS) examinations of suspicious breast masses in comparison to handheld breast ultrasound (HHUS) with regard to Breast Imaging Reporting and Data System (BI-RADS) category assessment, and to investigate the factors affecting discrepancies in categorization. A total of 135 masses that were assessed as BI-RADS categories 4 and 5 on ABUS and underwent ultrasound-guided core needle biopsy from May 2017 to December 2017 were included in this study. The BI-RADS categories were re-assessed using HHUS. Agreement of the BI-RADS categories was evaluated using kappa statistics, and the positive predictive value of each examination was calculated. They reported that the overall agreement between ABUS and HHUS in all cases was good (79.3%, kappa=0.61, P<0.001).

The authors concluded that the agreement between ABUS and HHUS examinations in the BI-RADS categorization of suspicious breast masses was good.

It is crucial to understand that for assessing the reliability of a qualitative variable, applying kappa statistics is not always appropriate. First, the use of kappa statistics depends on the prevalence in each category. Table 1 shows that in both situations (a) and (b), the prevalence of concordant cells is 90% and that of discordant cells is 10%; however, we get a different kappa value in each situation (0.44, interpreted as moderate, and 0.81, interpreted as very good, respectively). Second, whether it is appropriate to use kappa statistics also depends on the number of categories [2–7]. When a variable with more than two categories or an ordinal scale is used (with 3 or more ordered categories), then the weighted kappa would be a good choice (Table 2). Finally, another important flaw occurs when the two raters have unequal marginal distributions of their responses.

It is vitally important to recognize that reliability (precision, repeatability) and validity (accuracy) are completely different methodological issues. Positive predictive value is an estimate used to assess validity and has nothing to do with reliability. Other well-known estimates are sensitivity, specificity, negative predictive value, the positive likelihood ratio (ranging from 1 to infinity; the higher the positive likelihood ratio, the more accurate the test), and the negative likelihood ratio (ranging from 0 to 1; the lower the negative likelihood ratio, the more accurate the test) [8–10].

In this letter, I discussed some important limitations of applying kappa statistics to assess reliability.

Table 1. Limitation of the Cohen kappa for assessing reliability in two observers with different prevalence values for two categories

Observer 2	Observer 1		Total
	A	B	
Situation (a)			
A	85	5	90
B	5	5	10
Total, κ=0.44 (moderate)	90	10	100
Situation (b)			
A	45	5	50
B	5	45	50
Total, κ=0.81 (very good)	50	50	100

Values are presented as percentage.

Table 2. The kappa and weighted kappa values for calculating agreement between two observers for more than two categories

Observer 2	Observer 1			Total
	A	B	C	
A	60	20	1	81
B	2	12	4	18
C	3	11	11	25
Total	65	43	16	124
Estimate				
Kappa	0.43			
Weighted kappa	0.63			

Any conclusion regarding reliability needs to be supported in light of the methodological and statistical issues mentioned above. Otherwise, misinterpretations are inevitable.

ORCID: Siamak Sabour: https://orcid.org/0000-0002-1928-992X

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Yun G, Kim SM, Yun B, Ahn HS, Jang M. Reliability of automated versus handheld breast ultrasound examinations of suspicious breast masses. Ultrasonography 2019;38:264-271.
2. Szklo M, Nieto FJ. Epidemiology: beyond the basics. 3rd ed. Manhattan, NY: Jones and Bartlett Publisher, 2014.
3. Naderi M, Sabour S. Inter and intraobserver reliability and critical analysis of the FFP classification of osteoporotic pelvic ring injuries: Methodological issue. Injury 2019;50:1261-1262.
4. Sabour S. Methodologic concerns in reliability of noncalcified coronary artery plaque burden quantification. AJR Am J Roentgenol 2014;203:W343.
5. Sabour S, Dastjerdi EV. Reliability of four different computerized cephalometric analysis programs: a methodological error. Eur J Orthod 2013;35:848.
6. Sabour S. Reliability of immunocytochemistry and fluorescence in situ hybridization on fine-needle aspiration cytology samples of breast cancers: methodological issues. Diagn Cytopathol 2016;44:1128-1129.
7. Sabour S. Reliability of the ASA physical status scale in clinical practice: methodological issues. Br J Anaesth 2015;114:162-163.
8. Sabour S. A common mistake in assessing the diagnostic value of a test: failure to account for statistical and methodologic issues. J Nucl Med 2017;58:1182-1183.
9. Sabour S, Ghassemi F. The validity and reliability of a signal impact assessment tool: statistical issue to avoid misinterpretation. Pharmacoepidemiol Drug Saf 2016;25:1215-1216.
10. Sabour S. Validity and reliability of the new Canadian Nutrition Screening Tool in the 'real-world' hospital setting: methodological issues. Eur J Clin Nutr 2015;69:864.